Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML

Haïfa Zargayouna, Sylvie Salotti

Mesure de similarité dans une ontologie pour l'indexation sémantique de documents XML

Haïfa Zargayouna[1] et Sylvie Salotti[2]

[1] LIMSI/CNRS, Université Paris 11
haifa.zargayouna@limsi.fr
[2] LIPN - CNRS UMR 7030, Université Paris 13
sylvie.salotti@lipn.univ-paris13.fr

Résumé : Les documents XML posent de nouveaux défis et imposent de nouvelles méthodes de traitement d'information. Nous présentons dans cet article une mesure de similarité entre les concepts d'une ontologie que nous utilisons dans un système d'indexation de documents XML. Les documents sont structurés par un ensemble de balises sémantiquement pertinentes reliées à l'ontologie. Une partie des termes du corpus est également reliée à l'ontologie. Nous avons étendu le modèle vectoriel de Salton pour prendre en compte la structure des documents et le voisinage sémantique des termes.

Mots-clés : Similarité, ontologie, indexation sémantique.

1. Introduction

Nous présentons dans cet article une méthode d'évaluation de similarité mise en œuvre dans un système d'indexation sémantique de documents XML (documents textuels semi-structurés). L'avantage de ces documents est qu'ils possèdent une structure qui facilite leur présentation, ainsi que leur interprétation et leur exploitation dans des contextes présentant différents besoins. Cependant, très souvent, la majeure partie de l'information reste contenue dans les champs textuels, l'utilisation exclusive de la structure n'est donc pas suffisante. Nous proposons un système d'indexation permettant d'exploiter à la fois la structure et le contenu textuel des documents.

XML s'est imposé comme format standard de documents et un nombre de plus en plus important de documents sont disponibles en format XML. Cependant l'information apportée par les balises peut varier d'un simple découpage de la structure du document (titre, sections, paragraphe) à un véritable découpage sémantique dans lequel les balises donnent des informations sur le contenu des éléments textuels. De plus en plus de travaux visent à obtenir un tel balisage sémantique des documents, nous nous plaçons donc dans cette hypothèse. Par exemple un ensemble de compte-rendus médicaux pourront être structurés à l'aide de balises <info-patient>, <antécédent>, <traitement>… Cette structure nous permet de considérer le document comme un ensemble d'**unités sémantiques** représentant chacune un **contexte** particulier d'occurrence des termes. La prise en compte des unités sémantiques permet dans le cadre de Recherche d'Information (RI) de faire des recherches plus précises en indiquant l'unité d'information à rechercher et celle à retourner. Par exemple : rechercher les traitements préconisés pour tel type de symptôme. La structuration des textes peut être mise à profit dans le cadre de RàPC textuel (Textual CBR) où la phase de recherche peut être utilisée lors de la phase de remémoration des cas dans le cadre d'applications où le RàPC est utilisé pour des tâches d'aide au diagnostic ou à l'interprétation (Zargayouna & Salotti, 2004). Nous avons étendu le modèle vectoriel de Salton (Salton, 1971) (Salton & McGill, 1983) en effectuant le calcul du poids des termes pour chaque unité sémantique. Un document n'est donc plus représenté par un vecteur mais par un ensemble de vecteurs, chacun correspondant à une unité sémantique.

Par ailleurs, nous proposons d'utiliser une ontologie du domaine pour enrichir le calcul du poids des termes en intégrant la notion de voisinage sémantique. Les balises pertinentes et une partie des termes du corpus sont reliés à des concepts organisés dans une ontologie ou une taxonomie (dans un premier temps, nous n'avons utilisé que les liens de spécialisation/généralisation). La figure 1 présente la structure de l'index.

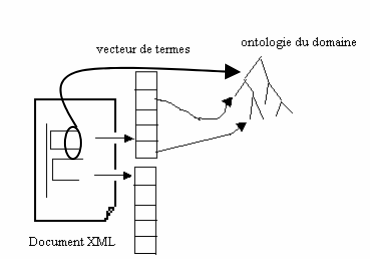


Fig. 1 – Structure de l'index

La similarité entre les termes des documents implique l'évaluation d'une similarité entre les concepts de l'ontologie. La détermination du degré de similarité entre deux concepts reliés à des termes d'un document est un problème qui se pose dans beaucoup d'applications : désambiguïsation, résumé automatique, extraction d'information, indexation automatique, recherche par similarité, etc. Nous présentons ici une mesure de similarité entre les concepts qui nous permet d'intégrer la notion de voisinage sémantique lors du calcul du poids des termes. L'utilisation de la structure du document nous permet de limiter l'étendue des calculs de similarité lors de la phase d'indexation. En effet, les calculs de similarité entre concepts sont faits en fonction de l'unité sémantique.

Dans la section suivante, nous présentons le processus général d'indexation qui mène au calcul de la similarité. (Zargayouna, 2004) présente une description détaillée du système. Nous présentons ensuite brièvement la problématique de l'utilisation d'ontologies, ou plus généralement de ressources sémantiques, dans les systèmes de recherche d'information. Nous décrivons ensuite en section 4 différentes mesures qui ont été définies pour évaluer la similarité entre les concepts d'une ontologie. En section 6, nous définissons la mesure que nous proposons d'utiliser. En section 5, nous présentons comment cette similarité entre concepts est utilisée dans l'évaluation de la similarité entre documents. Nous concluons en soulignant les avantages et les limites de notre approche et en discutant le problème de la validation d'un tel système d'indexation.

2. Processus général

Le cadre de ce travail est l'indexation de documents XML de spécialité pour lesquels nous disposons d'une ontologie du domaine qui peut être construite à partir des corpus ou en utilisant différentes ressources. Le choix de traiter des corpus spécialisés simplifie la tâche en limitant le vocabulaire, la polysémie et la variabilité des formes syntaxiques.

L'objectif de notre système est de construire un index qui prend en considération la structure et le contenu des documents. Le processus d'indexation comporte les phases suivantes (voir figure 4):

- Pour chaque document, nous extrayons la structure donnée par les balises par le parseur SAX (Simple API for XML). Nous modélisons la structure XML par un arbre étiqueté où chaque élément (ou attribut) correspond à un nœud. Nous ne faisons aucune distinction entre les éléments et les attributs. Nous calculons l'arbre minimal des structures présentes, où nous ne gardons que des chemins uniques.

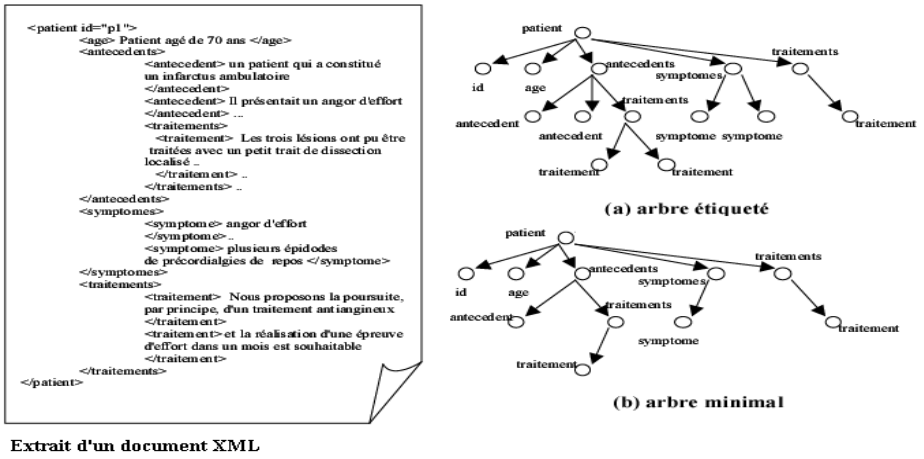


Fig. 2 – Exemple d'un document XML avec sa représentation en arbre et sa représentation réduite

- Chaque chemin unique représente une unité d'information ayant son contexte. Ces unités (qu'on appelle unités sémantiques) peuvent faire référence à des concepts dans l'ontologie. Nous les rattachons dans un premier temps manuellement. Ces concepts sont appelés **concept contexte** (voir figure 3), ils servent à retrouver des structures similaires sémantiquement et ayant des labels différents. Nous verrons dans ce qui suit leur importance dans le calcul de la similarité conceptuelle.

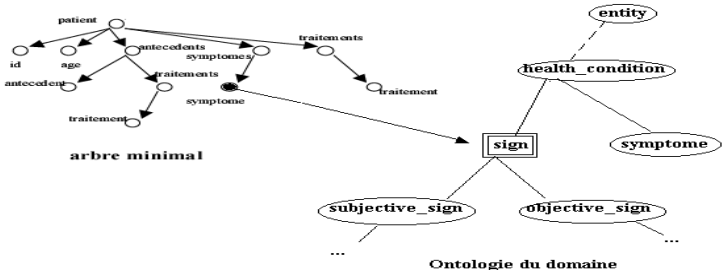


Fig. 3 – Rattachement d'unité sémantique au concept contexte

- Pour chaque unité sémantique, nous extrayons les candidats termes. Nous utilisons l'étiqueteur Treetagger qui fournit en sortie la catégorie grammaticale et le lemme de chaque terme. Seul un sous-ensemble des catégories grammaticales nous intéresse, il s'agit dans un premier temps des mots dits pleins : noms, verbes, adjectifs. Des erreurs d'étiquetage peuvent survenir mais par souci d'automatisation, nous n'avons pas d'étape de validation des candidats termes, ils sont directement intégrés dans l'index. C'est pour cela qu'on parlera de termes dans le reste de l'article.

- Nous calculons les fréquences pondérées des termes en fonction de leur nombre d'occurrence dans l'unité sémantique et dans le document.

- Nous rattachons le plus de termes possibles aux concepts de l'ontologie. Nous utilisons les formes lemmatisées des termes et les associons aux entrées lexicales des concepts. Cette phase n'est pas triviale, beaucoup de travaux s'y sont intéressés. (Volk et al. 2002), (Volk et al., 2003) extraient aussi la forme lemmatisée des termes et procèdent à une phase de filtrage et normalisation (enlever les termes longs, inverser les variantes des termes). Un outil d'annotation apparie les termes aux concepts et permet de retrouver des termes composés Si un terme a plusieurs sens, plusieurs annotations lui sont assignées qui correspondent aux différentes interprétations L'automatisation de cette tâche n'est pas évidente. Outre le traitement de l'ambiguïté qui peut être fait dans cette phase ou relégué à la phase d'utilisation de l'ontologie en contexte, d'autres problèmes persistent. En effet, des termes présents dans le corpus peuvent ne pas avoir de concordance avec les entrées lexicales des concepts, ce problème est soulevé dans (Simon et al., 2003) pour le traitement des « concepts inconnus » du thésaurus.

- Nous calculons les similarités entre les concepts relatifs aux termes avec ceux des autres termes co-occurrents dans la même unité sémantique. Nous enrichissons les fréquences des termes avec ces similarités.

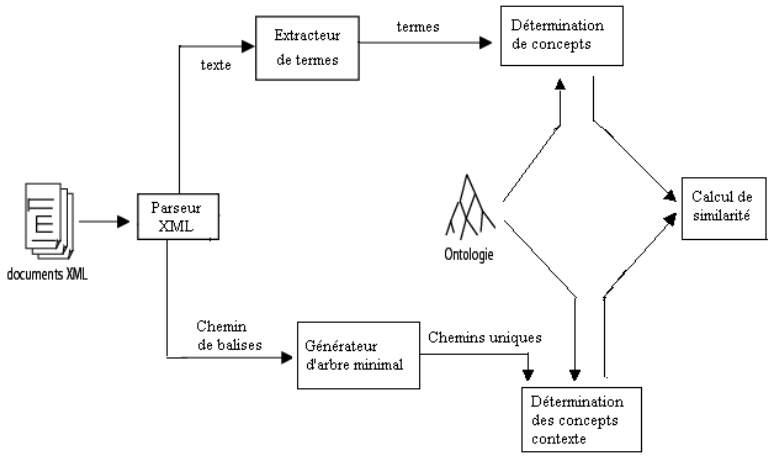


Fig. 4 – Processus d'indexation

Les termes qui ne sont pas rattachés à des concepts sont intégrés dans l'index. Ce qui nous permet de faire une recherche par concepts ainsi que par mots clés. Ceci est précieux dans le cas où des documents sont ajoutés à la base documentaire, même si leur contenu n'est pas rattaché à l'ontologie, nous pouvons quand même les retrouver.

3. Apport de la sémantique en Recherche d'Information

Les ressources sémantiques (thésaurus, ontologies, etc.) ont un apport considérable pour le traitement des documents textuels ou multimédia. Leur utilisation en Recherche d'Information (RI) peut intervenir lors de la phase de recherche ou lors de la phase d'indexation. La phase de recherche consiste à retrouver les documents les plus pertinents par rapport à une requête donnée. En général les documents retournés sont ordonnés à l'aide d'une mesure de similarité calculée entre le document et la requête. La phase d'indexation consiste à construire au préalable une structure d'accès aux documents qui facilitera la phase de recherche. Plus la phase d'indexation est sophistiquée, plus la phase de recherche sera facile.

3.1 Phase de recherche

L'intérêt d'utiliser des ressources sémantiques en recherche d'information est de pouvoir retourner, lors d'une recherche par similarité, les documents qui partagent avec la requête le maximum de concepts plutôt que le maximum de mots-clés. Les ontologies ont montré leur efficacité en RI (Andreasen et al. 2003), leur utilité s'est vu confirmé par le web sémantique. Une ontologie permet d'affiner les résultats en réduisant le silence et le bruit (Hernandez & Aussenac-Gilles, 2004). Les réseaux sémantiques ont montré leur apport en expansion de requêtes (Lu & Keefer, 1994) (Baziz et al., 2003). Le but de l'expansion de requête est soit d'élargir l'ensemble de documents retournés ou d'augmenter la précision. Dans le premier cas, la requête peut être étendue en ajoutant des termes similaires (généralement des synonymes) à ceux de la requête. Dans le deuxième cas, les termes peuvent être complètement changés pour reformuler la requête, une technique utilisée dans les retours arrière sur pertinence (Buckley et al., 1994).

3.2 Phase d'indexation

Les documents peuvent être indexés par un groupe de concepts, où on sait qu'un tel document traite des concepts A et B mais on ne connaît pas les relations entre eux dans le texte. Une autre méthode attribue à chaque document une description sémantique où les concepts sont représentés avec leurs relations sémantiques (Alhulou et al., 2003). Cette représentation confère un grand pouvoir d'expression mais peut par ce fait ralentir les traitements et la construction des descriptions sémantiques associées à chaque document n'est pas une tâche facile.

L'indexation automatique dans les deux cas pose des problèmes notamment celui de l'ambiguïté des termes (homonymie et polysémie) et on a généralement recours à des outils de Traitement Automatique des Langues (TAL). (Steffen et al., 2003) procèdent à une analyse lexicale: partie du discours, morphologie et analyse de phrases pour l'allemand. (Mihalcea & Moldovan, 1999) calculent la densité sémantique entre des couples de mots ce qui leur permet de retrouver le sens voulu des termes. (Stetina et al., 1998) travaillent sur des petits corpus d'apprentissage en considérant les phrases comme unité de contexte. Ils exploitent la probabilité d'apparition des termes dans ces unités ainsi que les liens sémantiques explicites. Mais ces techniques ne résolvent pas totalement le problème et il faut toujours faire le compromis entre la finesse des traitements et la complexité des systèmes. (Krovetz, 1997) a montré la nécessité d'indexer par les concepts (i.e. sens des mots) ainsi que les mots. Indexer les documents par les concepts uniquement peut induire en erreur car les techniques de désambiguïsation ne sont pas complètement fiables et se baser uniquement dessus risque d'entraîner une perte d'information. Nous indexons dans notre système les termes indépendamment du fait qu'ils soient reliés ou pas à une ontologie. Les liens sémantiques constituent ainsi un plus mais un terme qui n'est pas relié à l'ontologie peut aussi être retrouvé. Nous montrerons aussi que le problème d'ambiguïté est pris en charge en intégrant la notion de contexte dans nos calculs de similarité.

4. Mesures de similarité entre concepts

Rada et al. (Rada et al., 1989) ont suggéré que la similarité dans un réseau sémantique peut être calculée en se basant sur les liens taxonomiques « is-a ». Plus généralement, le calcul de la similarité entre concepts peut être basé sur les liens hiérarchiques de spécialisation/généralisation. Un moyen des plus évidents pour évaluer la similarité sémantique dans une taxonomie est alors de calculer la distance entre les concepts par le chemin le plus court. Les auteurs soulignent que cette proposition est valable pour tous les liens de type hiérarchique (is-a, kind-of, part-of, …) mais doit être adaptée pour d'autres types de liens (cause, etc.).

(Budanitisky & Hirst, 2001) comparent cinq mesures de similarités ou distances sémantiques utilisant WordNet (Fellbaum, 1998) (où la relation « is-a » est restreinte aux noms et verbes). Nous présentons quelques unes de ces mesures, un état de l'art complet est présenté par

(Patwardham, 2003) où sont comparées ces différentes mesures entre elles par rapport à des évaluations faites par des sujets humains.

Les mesures varient du simple calcul du nombre d'arcs à l'intégration de mesures statistiques.

(Hirst & St Onge, 1998) calculent la proximité sémantique (*semantic relatedness*) qui est une notion plus large que la similarité sémantique. Toutes les relations dans WordNet sont prises en compte. Les liens sont classés comme *haut* (eg. partie-de), *bas* (eg. sous-classe), *horizontal* (eg. antonyme). La relation est calculée entre termes par le poids du chemin le plus court qui mène du synset du terme à un autre. Il est calculé en fonctions de ces classifications qui indiquent les changements de direction :

$$Rel(c1, c2) = T - chemin - k \times d \quad (1)$$

Tels que *T* et *K* sont des constantes, *chemin* est la longueur du chemin le plus court en nombre d'arcs et *d* est le nombre de changements de direction.

L'idée est que deux termes sont proches sémantiquement si leurs synsets sont connectés par un chemin qui n'est pas très long et qui ne change pas souvent de direction. S'il n'y a pas de chemin, le calcul est égal à zéro.

Cette mesure s'éloigne de la similarité proprement dite car elle traite tout type de relations, comparée aux autres mesures de similarité elle ne donne pas, de ce fait, de bons résultats.

Les mesures de (Resnik, 1995) et (Jiang & Conrath, 1997) sont fondées sur la notion de contenu informationnel. La notion de contenu informationnel (CI) a été la première fois introduite par Resnik. Elle utilise conjointement l'ontologie et le corpus. Le contenu informationnel d'un concept traduit la pertinence d'un concept dans le corpus en tenant compte de la fréquence de son apparition dans le corpus ainsi que de la fréquence d'apparition des concepts qu'il subsume. On dit qu'un concept C1 *subsume* un concept C2 si C2 est plus spécifique que C1. Plus précisément le contenu informationnel se calcule de la manière par la formule suivante :

$$CI(c) = - \log(P(c)) \quad (2)$$

Où P(c) est la probabilité de retrouver une instance du concept c. Ces probabilités sont calculés par : frequence(c)/N où N est le nombre total de concepts. Voici un extrait de WordNet, le nombre attaché à chaque nœud est P(c) (Lin, 1998).

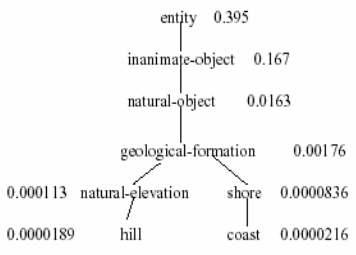


Fig. 5 – Extrait de Wordnet

Resnik définit la similarité sémantique entre deux concepts par la quantité d'information qu'ils partagent. Cette information partagée est égale au contenu informationnel du plus petit généralisant (PPG) – le concept le plus spécifique qui subsume les deux concepts dans l'ontologie.

$$Sim(c1, c2) = CI\ (ppg(c1, c2)) \quad (3)$$

Cette mesure ne dépend que du PPG et est de ce fait un peu sommaire car nous pouvons avoir ppg(a,b) = ppg(d,e) même si d et e sont plus proches du PPG que a et b.

La mesure de (Jiang & Conrath, 1997) pallie aux limites de la mesure de Resnik en combinant le contenu informationnel du PPG à ceux des concepts. Elle prend en considération aussi le nombre d'arcs. Ainsi une distance est définie :

$$distance(c1, c2) = CI(c1) + CI(c2) - (2.CI(ppg(c1, c2))) \quad (4)$$

La mesure de similarité devient donc :

$$Sim(c1, c2) = 1/distance(c1, c2) \quad (5)$$

(Wu & Palmer, 1994) ont défini une mesure de similarité entre concepts pour la traduction automatique entre l'anglais et le chinois. Leur mesure s'applique à un domaine conceptuel qui correspond à un point de vue donné. La similarité est définie par rapport à la distance qui sépare deux concepts par rapport à leur PPG ainsi que la racine de la hiérarchie. La similarité entre C1 et C2 (voir figure 6) est :

$$ConSim(C1, C2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \quad (6)$$

Plus formellement cette mesure devient :

$$ConSim(C1, C2) = \frac{2*depth(C)}{depth_C(C1)+depth_C(C2)} \quad (7)$$

Où C est le PPG de C1 et C2 (en nombre d'arcs), depth(C) est le nombre d'arcs qui séparent C de la racine et $depth_c(C_i)$ avec i le nombre d'arcs qui séparent C_i de la racine en passant par C.

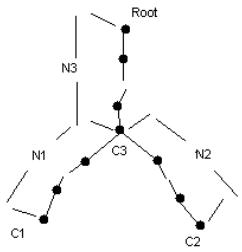


Fig. 6 – Les relations conceptuelles (Wu & Palmer, 1994)

Cette mesure a l'avantage d'être simple à implémenter et d'avoir d'aussi bonnes performances que les autres mesures de similarité (Lin, 1998).

5. Notre mesure

Dans un précédent travail (Zargayouna, 2001) dans le cadre d'une RI multimédia, nous avons calculé les similarités entre des descriptions formalisées en logique de description. La similarité entre deux concepts est le PPG les subsumant. Nous calculons la description symbolique de ce PPG, il peut de ce fait ne pas exister dans l'ontologie. Ceci est intéressant car nous arrivons ainsi à déterminer des similarités à une granularité plus fine que celle de l'ontologie. De plus nous utilisons toutes les relations qui existent entre les concepts. La contrepartie d'une telle précision est

la complexité des calculs qui reste quand même considérable. Une des limites de ce travail est le manque de relation d'ordre total entre les similarités. Ce problème peut être résolu par les mesures numériques de calcul de la similarité conceptuelle. Nous l'appliquons aux données textuelles. Nous nous inspirons de la mesure de (Wu & Palmer, 1994) présentée ci-dessus. Nous n'utilisons pas la notion de contenu informationnel car elle serait redondante puisque nous combinons la mesure de similarité à la mesure distributionnelle (TF-ITDF) des termes dans les documents. La mesure de (Wu & Palmer, 1994) a été utilisé par (Halkidi et al., 2003) pour organiser des documents web dans des clusters. Elle a aussi servi dans (Desmontils & Jacquin ,2001) pour évaluer la proximité sémantique de deux concepts d'une page html relativement à un thésaurus dans le cadre d'une indexation d'un site web par des ontologies.

La mesure de (Wu & Palmer, 1994) est intéressante mais présente une limite car elle vise essentiellement à détecter la similarité entre deux concepts par rapport à leur distance de leur PPG. Plus ce subsumant est général, moins ils sont similaires (et inversement). Cependant, elle ne capte pas les mêmes similarités que la similarité conceptuelle symbolique. Ainsi on peut avoir conSim(A, f) < conSim(A,B), f étant un des fils de A et B un des frères de A. Ce qui est à notre sens inadéquat dans le cadre de recherche d'information où il faut ramener tous les fils d'un concept (i.e requête) avant son voisinage.

Nous définissons *spec(C1, C2)* une fonction qui calcule la spécificité de deux concepts par rapport au concept le plus bas de l'ontologie (bottom, concept virtuel qui symbolise la fin de l'ontologie) comme le montre la figure 7. Cette fonction servira à pénaliser les concepts qui ne sont pas dans la même lignée. Ainsi on s'assure que les fils sont pris en compte en priorité et qu'aucun concept du voisinage ne sera plus similaire que les fils.

spec(C1, C2) = N4*N1*N2. (voir figure 7) Plus formellement :

$$spec(C1, C2) = depth_b(C) * distance(C,C_1) * distance(C,C_2) \quad (8)$$

avec $depth_b(C)$ est le nombre maximum d'arcs qui séparent C de *bottom* et $distance(C, C_i)$ la distance en nombre d'arcs entre C et C_i

spec(C1, C2) est nulle si C1 est ancêtre de C2 ou l'inverse. Seront pénalisés donc les concepts voisins de C1 ou C2.

Ainsi la mesure de similarité (équation (7)) devient :

$$sim(C1, C2) = \frac{2*depth(C)}{depth_C(C1)+depth_C(C2)+spec(C1,C2)} \quad (9)$$

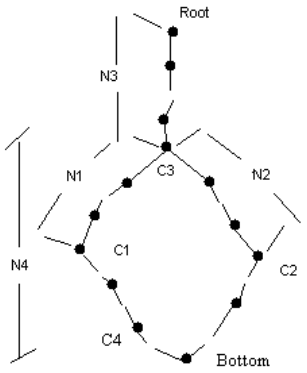
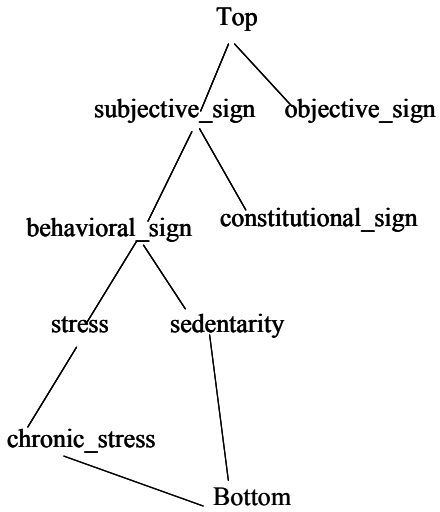


Fig. 7 – Les nouvelles relations conceptuelles

Dans l'exemple suivant, nous présentons un extrait de l'ontologie Menelas ainsi que les calculs de similarités (ConSim pour la mesure de (Wu & Palmer, 1994) et sim pour notre mesure). La similarité entre behavioral_sign et constitutional_sign (lien entre frères) se trouve réduite par notre mesure, celle de behavioral_sign et chronic_stress (lien père/fils) reste inchangée. Nous nous assurons en calculant la distance par rapport à bottom que sim(behavioral_sign,constitutional_sign) > sim(behavioral_sign, F), tel que F ∈ ensemble des fils de behavioral_sign.



ConSim(behavioral_sign, constitutional_sign) = 2*1/(2+1+1)= 0.5

ConSim(behavioral_sign, chronic_stress) = 2*1/(2+2+0)= 0.5

Sim(behavioral_sign, constitutional_sign) = 2*1/(2+1+1)+(4*1*1)= 0.25

Sim(behavioral_sign, chronic_stress) = 2*1/(2+2+0)+(3*0*2)= 0.5

6. Similarité entre documents

Les documents sont représentés par des ensembles de vecteurs de termes. Chaque unité de contexte génère un vecteur. Les poids des termes sont calculés en fonction de leur distribution dans les balises. Le poids d'un terme est enrichi par les similarités conceptuelles des termes co-occurrents dans la même balise. Il est calculé pour un document et un contexte (à savoir la balise) donnés.

Ce poids noté SemW(t,b,d) est calculé de la manière suivante :

$$SemW(t,b,d) = TF\text{-}ITDF(t,b,d) + \left(\sum_{i=1}^{n} Sim(t,t_i) * TF\text{-}ITDF(t_i,b,d)\right)/n \qquad (10)$$

avec $Sim(t,t_i) > \textbf{seuil}$; n le nombre de termes dans la balise b et seuil une valeur qui fixe la similarité à un certain voisinage, nous la fixons dans un premier temps à la similarité entre le concept de t et le **concept contexte** (concept qui représente la balise). TF-ITDF (Term Frequency–Inverse Tag and Document Frequency) est le poids initial attribué aux termes en fonction du document et de la balise dans lesquels ils apparaissent.

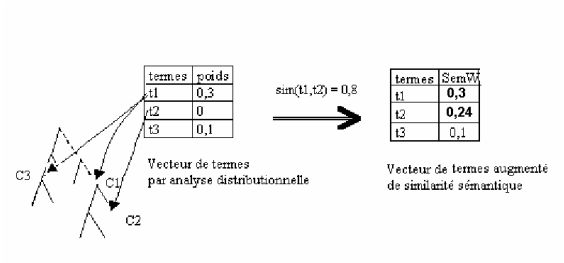


Fig. 8 – Prise en compte de la similarité sémantique

Le terme t2 qui avait une pondération nulle se trouve enrichi par le poids sémantique de t1 qui lui est proche dans l'ontologie, t2 peut de ce fait être retrouvé lors de la phase de recherche. On remarque que t1 ne se trouve pas enrichi par le poids de t2 ce dernier étant absent dans le document.

Le calcul de la similarité entre les termes co-occurrents dans la même balise nous permet de gérer aussi en partie le problème d'ambiguïté sémantique. En effet, dans la figure 8 le terme t1 est rattaché à deux concepts différents. Sim(t1, t2) est égal à la somme de sim(C1, C2) et sim(C3,C2), C3 étant loin sémantiquement de C2, il ne sera pas pris en considération et C1 se trouve enrichi seulement par le poids de C2. Il est à noter que le poids de t3 reste inchangé du fait qu'il n'est rattaché à aucun concept. Ceci est très important car nous permet de faire une recherche par concepts ainsi que simplement par mots clés.

Les similarités entre vecteurs sont calculés par leur cosinus :

$$\text{cosinus}(V1,V2)= \frac{\sum_{i=1}^{m} v1_i.v2_i}{\sqrt{\sum_{i=1}^{m} v1_i.v1_i} * \sqrt{\sum_{i=1}^{m} v2_i.v2_i}} \qquad (11)$$

Où $v1i$ et $v2i$ représentent les éléments des vecteurs V1 et V2 de taille m.

La similarité entre les documents est calculée par une agrégation des similarités entre les vecteurs.

7. Conclusion

Pour intégrer la notion de voisinage sémantique, nous avons utilisé une ontologie de concepts auxquels sont reliés les termes des documents. Dans un premier temps nous n'avons pris en considération que les liens de spécialisation/généralisation entre les concepts. En nous basant sur la mesure de similarité entre concepts présentée par (Wu & Palmer, 1994), nous avons proposé une nouvelle mesure telle que les descendants directs d'un concept sont considérés plus similaires au concept que ses frères. Nous avons alors défini un nouveau calcul du poids des termes SemW qui tient compte de la similarité conceptuelle entre les termes du même contexte. Le calcul de la similarité sémantique lors de l'indexation allége les traitements lors de la recherche.

Une des limites de notre approche, tient au fait que nous supposons disposer d'une ontologie de concepts reliée au corpus. Rappelons que nous nous plaçons dans le cadre de l'indexation de documents structurés, pour lesquels on peut supposer qu'il existe certaines ressources sur le vocabulaire du domaine. Pour utiliser le modèle présenté dans cet article, il suffit de disposer d'une structure hiérarchique entre concepts correspondant aux liens de spécialisation/généralisation. Cependant, nous sommes conscientes que le calcul de la mesure de similarité par restriction sur le lien « is-a » n'est pas toujours bien adapté, les autres types de liens peuvent être aussi importants dans le calcul de la similarité. Nous envisageons de travailler sur la

prise en compte d'autres types de liens comme par exemple le lien de composition. De plus, dans la réalité, les taxonomies ne sont pas toujours au même niveau de granularité, des parties peuvent être plus denses que d'autres. Ces problèmes peuvent être résolus, en partie, en associant des poids aux liens. L'affectation de ces poids peut être basée sur : les types de liens présents, la profondeur du lien dans la taxonomie et la densité du concept par ses voisins immédiats.

L'évaluation de notre mesure de similarité est nécessaire pour tester son efficacité ainsi que la pertinence d'un tel calcul lors de la phase d'indexation. Trois approches existent pour tester l'efficacité des mesures de similarité (Budanitisky & Hirst, 2001): la première étudie le cadre théorique de telles mesures par leurs propriétés et les cas qu'ils traitent, etc. Une deuxième manière consiste à comparer ces mesures par rapport à un jugement humain mais il est difficile de mettre en place de telles expérimentations qui porteraient sur un ensemble assez significatif de concepts. La troisième approche compare ces mesures par rapport à leur performance dans un cadre particulier d'une application TAL. Dans (Budanitisky & Hirst, 2001) cette application consiste à détecter et corriger des mots mal orthographiés.

Nous pouvons évaluer directement la structure d'index. Il s'agit généralement de calculer le temps d'indexation, l'espace de stockage de l'index par rapport à la taille de la base documentaire. Comme nous utilisons une ontologie, sa construction et son rattachement au corpus font partie de la phase d'indexation. Le calcul du temps de construction de l'index ne permet pas de juger de la valeur de l'index.

On peut aussi évaluer la pertinence d'un index en testant son impact sur la recherche, en utilisant les mesures de pertinence classiques de rappel et précision ou l'exhaustivité et la pertinence. La difficulté de l'évaluation de notre système est d'avoir un corpus avec des balises XML «pertinentes» (en vue d'une recherche structurée) et une ontologie associée. Quand l'ontologie est créée à partir du corpus manuellement ou par des méthodes semi-automatiques (Szullman et al. 2002), le lien entre les termes et le concept est évident. Le problème se pose quand on dispose d'un corpus de spécialité et d'une ontologie du domaine, l'appariement entre terme et concept n'est pas toujours évident.

Références

Alhulou R., Napoli A. & Nauer E. (2003). Une mesure de similarité pour raisonner sur des documents, Actes des Journées Nationales sur les Modèles de Raisonnement, Paris, 27-28 novembre 2003.

Andreasen T., Bulskov H. & Knappe R. (2003) On Ontology-based Querying, in *18th International Joint Conference on Artificial Intelligence, Ontologies and Distributed Systems, IJCAI 2003*

Szullman S., B.Biebow & Aussenac-Gilles N., « Structuration de terminologies à l'aide d'outils de TAL avec TERMINAE », *Revue Traitement Automatique des Langues*, vol. 43, 2002.

Baziz M., Aussenac-Gilles N. & Boughanem M. (2003) Exploitation des Liens Sémantiques pour l'Expansion de Requêtes dans un Système de Recherche d'Information. Dans : *XXIème Congrès INFORSID 2003*

Buckley, C., Salton, Allan, G., J. & Singhal, A. (1994). Automatic query expansion using SMART: TREC 3. In *Proceedings of TREC-3*.

Budanitsky, A. & Hirst, G. (1998). Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measures. In *Workshop on WordNet and Other Lexical Resources, in the North American Chapter of the Association for Computational Linguistics (NAACL-2001*), Pittsburgh, PA.

Desmontils E. & Jacquin C. (2001). Des ontologies pour indexer un site Web. *Dans actes des journées francophones d'Ingénierie des Connaissances. IC' 2001.*

Fellbaum, C. (1998). WORDNET. An Electronic Lexical Database. In *The MIT Press.*

Halkidi M. & Nguyen B & Varlamis I. & Vazirgiannis M. (2003) Thesus: Organising Web Document Collections based on Semantics and Clustering, *Journal on Very Large Databases, Special Edition on the Semantic Web, Novembre 2003*

Hernandez N. & Aussenac-Gilles N. (2004). OntoExplo : Ontologies pour l'aide à une activité de veille ou d'exploration d'un domaine. dans *VIème Journées de l'innovation, Foix, 28-29 Janvier 2004*

Hirst G. & St Onge D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. In *Christiane Fellbaum (editor), WordNet: An electronic lexical database, Cambridge, MA: The MIT Press* .

Jiang J. & Conrath D. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings on International Conference on Research in Computational Linguistics*, Taiwan, 1997.

Krovetz R. (1997) Homonymy and polysemy in Information Retrieval. In Proceedings of ACL/EACL'97.

D. Lin. (1998) An information-theoretic definition of similarity. In *Proceedings of 15th International Conference On Machine Learning, 1998.*

Lu, X. A. & Keefer, R. B. (1994). Query expansion/reduction and its impact on retrieval effectiveness. *Overview of the Third Text REtrieval Conference (TREC-3), NIST Special Publication 500-225*, edited by D. K. Harman, 231-240.

Mihalcea, R. & Moldovan, D. (1999) A method for Word Sense Disambiguation of unrestricted text. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*

Patwardham S. (2003). Incorporating Dictionary and Corpus Information in a Measure of Semantic Relatedness, *M.S. Thesis*, August.

Rada R., Mili H., Bicknell E., & Blettner M. (1989). Development and application of a metric on semantic nets. *IEEE Transaction on Systems, Man, and Cybernetics*, 19(1):17--30.

Resnik P. (1995). Using information content to evaluate semantic similarity in a taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Montreal.

Salton G. (1971). The SMART Retrieval System – experiments. *in automatic document processing. U Perntice-Hall, Inc., Englewood Cliffs, NJ.*

Salton G. & McGill, M.J. (1983) Introduction to Modern Information Retrieval. *McGraw-Hill, New York.*

Simon L., Desmontils E. & Jacquin C. (2003) Utilisation de techniques d'enrichissement d'ontologie pour améliorer le processus d'indexation structurée dans *journées francophones d'ingénierie des connaissances (IC'2003)*

Steffen D., Sacaleanu B. & Paul Buitelaar. (2003) Domain Specific Sense Disambiguation with Unsupervised Methods ***In:*** *proceedings 1. GermaNet-Workshops des GLDV-AK Lexikografie.*

Stetina, J., Kurohashi, S. & Nagao, M. (1998). General word sense disambiguation method based on a full sentential context. *Usage of WordNet in Natural Language Processing, Proceedings of COLING-ACL Workshop, Montreal, Canada, July 1998*

Volk M., Vintar S. & Buitelaar P. (2003) Ontologies in Cross-Language Information Retrieval. In: *Proc. of 2nd Conference on Professional Knowledge Management*. Lucerne. 2003.

Volk M., Ripplinger B., Vintar Š., Buitelaar P., Raileanu, D. & Sacaleanu B. (2002) Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval *In:* ***International Journal of Medical Informatics****, Volume 67:1-3, December 2002.*

Wu Z. & Palmer M. (1994). Verb Semantics and Lexical Selection, *Proceedings of the 32nd Annual Meetings of the Associations for Computational Linguistics*, pages 133-138.

Zargayouna H. (2001). Raisonnement par similarité pour l'indexation et la recherche dans des documents multimédia. dans *Rapport interne LIMSI, N° 2001-12*, Juin 2001.

Zargayouna H. (2004). Contexte et sémantique pour une indexation de documents semi-structurés. À paraître *dans ACM COnférence en Recherche Information et Applications*, CORIA'2004.

Zargayouna H & Salotti S. (2004) Mesure de similarité sémantique pour l'indexation de documents semi-structurés dans *12ème Atelier de Raisonnement à Partir de Cas, Mars 2004*